Forgetting and stability of Score-based Generative Models.

Stanislas Strasman

June 19, 2026

Sorbonne University, Laboratoire de Probabilités, Statistique et Modélisation (LPSM)

On Forgetting and Stability of Score-Based Generative Models

Stanislas Strasman, Gabriel Victorino Cardoso, Sylvain Le Corff, Vincent Lemaire, Antonio Ocello

arXiv:2601.21868

Generative AI

Data

We observe a **finite** dataset

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \pi_{\text{data}}\,, \qquad \pi_{\text{data}} \in \mathcal{P}(\mathbb{R}^d) \quad \text{unknown}\,.$$

Unsupervised Machine Learning

Learn a sampling mechanism whose output law $\widehat{\pi}$ satisfies

$$\widehat{\pi} \approx \pi_{\text{data}}\,.$$

⚠ both a learning and a sampling problem.

What does $\approx$ mean?

The notion of approximation depends on the application: visual quality, downstream performance, or proper mathematical distance on the space of probability measures.

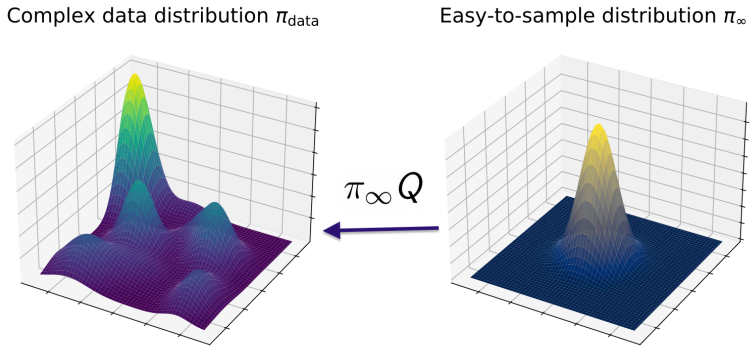
Generative Modeling design and stability

Simuable generation model

Choose an easy-to-sample prior π_∞ and a simulable transition kernel Q such that

$$\widehat{\pi} = \pi_\infty Q \qquad \text{and} \qquad \pi_\infty Q \approx \pi_{\text{data}}.$$

What can be said about the properties of Q ?



Roadmap

Forgetting in SGMs

- Mathematical framework.
- The log concave example
- Harris theory for Markov chains

Stability bound

- Presentation and usual approaches
- A forgetting-based approach

Numerical illustration

Forgetting in SGMs

Time Reversal of diffusion processes

Forward process

Forward process $(\overrightarrow{X}_t)_{t\in[0,T]}$ is solution to $\mathrm{d}\overrightarrow{X}_t = \sqrt{2}\,\mathrm{d}B_t, \quad \overrightarrow{X}_0 \sim \pi_{\mathrm{data}}.$

Backward process (Haussmann and Pardoux, 1986)

Time-reversed process for $T > 0$ fixed

$$(\overleftarrow{X}_t)_{t\in[0,T]} \overset{\mathcal{L}}{=} (\overrightarrow{X}_{T-t})_{t\in[0,T]}$$

with p_t the marignal p.d.f of the forward, satisfies

$$\mathrm{d}\overleftarrow{X}_t = 2\nabla \log p_{T-t}(\overleftarrow{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad \overleftarrow{X}_0 \sim p_T.$$

Backward process semigroup

For $0 \le s < t \le T$, and f measurable bounded $Q_{t|s}f(x) := \mathbb{E}\left[f(\overleftarrow{X}_t) \,\middle|\, \overleftarrow{X}_s = x\right]$, the ideal sampler satisfies

$$\boxed{\pi_{\mathrm{data}} = p_T Q_{T|0}}$$

What do we mean by forgetting?

$$Q_{t|s} f(x) = \mathbb{E}\left[f(\overleftarrow{X}_t) \middle| \overleftarrow{X}_s = x \right], \qquad 0 \le s < t \le T.$$

Forgetting of initial conditions

A distance ρ on probability measures exhibits forgetting if

$$\rho\big(\mu Q_{t|s}, \nu Q_{t|s}\big) \le \alpha_{s,t} \rho(\mu, \nu), \qquad \alpha_{s,t} < 1.$$

$\Longrightarrow$ μ and ν become closer after being pushed through the same backward dynamics.

Why does forgetting matter?

- **Robustness, stability and error propagation:** local errors damped exponentially fast by the remaining flow, not accumulated over time.
- **Dynamical interpretation.** mode or class selection.

What information is forgotten, and what information is preserved?

Related statistical physics viewpoints

Physics-inspired analyses

Study of dynamical regimes in the backward process includes phase transitions, symmetry breaking, speciation and collapse phenomena in diffusion models Ambrogioni (2024); Biroli et al. (2024).

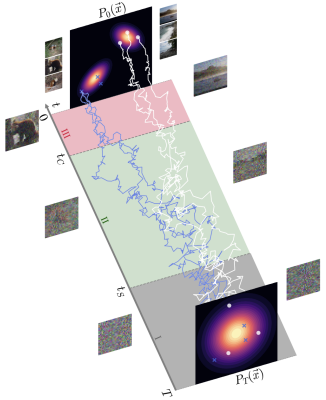


Illustration of speciation dynamics from Biroli et al. (2024).

Study the exact reverse dynamics under synchronous coupling

$$\mathrm{d}\overleftarrow{X}_t = 2\nabla \log p_{T-t}(\overleftarrow{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t. \tag{1}$$

Let $\overleftarrow{X}_t^x$ and $\overleftarrow{X}_t^y$ solve (1) with the same Brownian motion, starting from $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^d$. Set

$$Z_t := \overleftarrow{X}_t^x - \overleftarrow{X}_t^y.$$

$$\frac{\mathrm{d}}{\mathrm{d}t} Z_t = 2\left[\nabla \log p_{T-t}(\overleftarrow{X}_t^x) - \nabla \log p_{T-t}(\overleftarrow{X}_t^y)\right].$$

Hence

$$\frac{\mathrm{d}}{\mathrm{d}t}\|Z_t\|^2 = 4\left\langle Z_t, \nabla \log p_{T-t}(\overleftarrow{X}_t^x) - \nabla \log p_{T-t}(\overleftarrow{X}_t^y)\right\rangle.$$

The log-concave case II

Assume that, for every $s \in [0,t]$, p_{T-s} is λ_s-log-concave, i.e.

$$\langle x-y, \nabla \log p_{T-s}(x) - \nabla \log p_{T-s}(y)\rangle \le -\lambda_s \|x-y\|^2, \qquad \lambda_s > 0.$$

Then the synchronous coupling satisfies $\frac{\mathrm{d}}{\mathrm{d}s}\|Z_s\|^2 \le -4\lambda_s \|Z_s\|^2$. By Grönwall,

$$\|\overleftarrow{X}_t^x - \overleftarrow{X}_t^y\|^2 = \|Z_t\|^2 \le \exp\left(-4\int_0^t \lambda_s \,\mathrm{d}s\right)\|Z_0\|^2 = \exp\left(-4\int_0^t \lambda_s \,\mathrm{d}s\right)\|x-y\|^2 .$$

Wasserstein contraction

Taking the infimum over initial distribution gives

$$\mathcal{W}_2(\mu Q_{0,t}, \nu Q_{0,t}) \le \underbrace{\exp\left(-2\int_0^t \lambda_s \,\mathrm{d}s\right)}_{\alpha_{0,t}} \mathcal{W}_2(\mu,\nu)$$

Brownian Smoothing Preserves Strong Log-Concavity

Preservation under Gaussian convolution

If the data density is strongly log-concave, for every $t \geq 0$, p_t remains strongly log-concave:

$$\nabla^2 \log p_t(x) \preceq -\lambda_t \mathrm{I}_d, \qquad \lambda_t = \left(\frac{1}{\lambda_0} + 2t\right)^{-1}.$$

Too restrictive in practice

$$p(x) \propto e^{-U(x)}, \qquad \nabla^2 U(x) \succeq \lambda \mathrm{I}_d.$$

Thus U has one uniformly convex basin. Forgetting is almost built in: synchronous trajectories contract. But this excludes separated modes and many realistic data distributions.

Compared with curvature-based forgetting

Gentiloni-Silveri and Ocello (2025) viewpoint

OU noising can regularize the data law:

$$\text{weak log-concavity} \quad \rightsquigarrow \quad \text{log-concavity after enough noising.}$$

This yields contractive and non-contractive regimes for the reverse dynamics

$$t \geq t_\star \qquad \Longrightarrow \qquad \nabla^2 \log p_t(x) \preceq -\lambda_t \mathrm{I}_d, \qquad \lambda_t > 0.$$

Consequence for the reverse dynamics

On time intervals where p_{T-s} is log-concave, synchronous coupling gives a contractive regime:

$$\mathcal{W}_2(\mu Q_{s,t}, \nu Q_{s,t}) \leq \rho_{s,t}\, \mathcal{W}_2(\mu, \nu), \qquad \rho_{s,t} < 1.$$

Hence perturbations made before this contractive phase are partially forgotten.

⚠ Noising may create a forgetting window, even if the data law is not globally log-concave.

From curvature to Markov-kernel forgetting

What we have seen

Log-concavity gives a orgetting mechanism in Wasserstein 2:

$$\nabla^2 \log p_{T-s} \preceq -\lambda_s \mathrm{I}_d \quad \Longrightarrow \quad \mathcal{W}_2(\mu Q_{t|s}, \nu Q_{t|s}) \leq \alpha_{s,t} \mathcal{W}_2(\mu, \nu).$$

$\Longrightarrow$ synchronous coupling + curvature.

$\mathcal{W}_2$ analyses

In many $\mathcal{W}_2$-based convergence proofs, such contraction factors appear as a *stability tool*. Forgetting is usually not isolated as the main object of study. It is often a *by-product of the error analysis*.

Forgetting is aproperty of the Markov kernel

The object of interest is the reverse Markov kernel $Q_{t|s}$. Forgetting means a contraction property of the form

$$\rho(\mu Q_{t|s}, \nu Q_{t|s}) \leq \alpha_{s,t} \rho(\mu, \nu), \qquad \alpha_{s,t} < 1.$$

This property may hold even when there is no global convex geometry.

A strong route to forgetting: global Doeblin I

Global Doeblin condition

Let Q be a Markov kernel on E. Assume that there exist $\varepsilon \in (0,1]$ and a probability measure $\nu_\star$ such that

$$Q(x,\cdot) \geq \varepsilon \nu_\star(\cdot), \qquad \forall x \in E.$$

Mixture decomposition

If $\varepsilon < 1$, define

$$R(x,\cdot) := \frac{Q(x,\cdot) - \varepsilon \nu_\star(\cdot)}{1-\varepsilon}.$$

By minorization, the denominator is a positive measure and $R(x,E) = 1$. Then R is a Markov kernel and

$$Q(x,\cdot) = \varepsilon \nu_\star(\cdot) + (1-\varepsilon) R(x,\cdot).$$

Interpretation

With probability at least ε, the chain is refreshed from the same law $\nu_\star$, independently of its starting point.

A strong route to forgetting: global Doeblin II

For any two probability measures $\mu, \tilde{\mu}$, the decomposition gives

$$\mu Q = \varepsilon \nu_\star + (1-\varepsilon)\mu R, \qquad \tilde{\mu} Q = \varepsilon \nu_\star + (1-\varepsilon)\tilde{\mu} R.$$

Hence the common part cancels:

$$\mu Q - \tilde{\mu} Q = (1-\varepsilon)(\mu R - \tilde{\mu} R).$$

Therefore

$$\|\mu Q - \tilde{\mu} Q\|_{\mathrm{TV}} = (1-\varepsilon)\|\mu R - \tilde{\mu} R\|_{\mathrm{TV}} \leq (1-\varepsilon)\|\mu - \tilde{\mu}\|_{\mathrm{TV}}.$$

Iterating the contraction, gives geometric discounting

$$\boxed{\|\mu Q^n - \tilde{\mu} Q^n\|_{\mathrm{TV}} \leq (1-\varepsilon)^n \|\mu - \tilde{\mu}\|_{\mathrm{TV}}.}$$

Harris assumptions for one Markov kernel

Global Doeblin too strong on $\mathbb{R}^d$

A global Doeblin condition asks for $Q(x,\cdot) \geq \varepsilon \nu_\star(\cdot)$, $\forall x \in \mathbb{R}^d$.

This means that every $Q(x,\cdot)$, even when x is far away, must contain the same fixed amount of mass $\varepsilon \nu_\star$.

Gaussian data example

If $\pi_{\text{data}} = \mathcal{N}(0, \sigma_0^2)$, then

$$Q_{t|s}(x,\cdot) = \mathcal{L}(\overleftarrow{X}_t \mid \overleftarrow{X}_s = x) = \mathcal{N}(m_{s,t}x, \Gamma_{s,t}),$$

with

$$m_{s,t} = \frac{\sigma_0^2 + 2(T-t)}{\sigma_0^2 + 2(T-s)} > 0.$$

Thus, for every fixed $R > 0$,

$$Q_{t|s}(x, [-R, R]) \to 0 \qquad \text{as } |x| \to \infty.$$

⚠ No fixed positive common mass can be shared by all $Q_{t|s}(x,\cdot)$.

Harris replacement: local minorization

Global Doeblin is too strong, but a local version can hold.

Let Q be a Markov transition kernel on $\mathbb{R}^d$, and let $V : \mathbb{R}^d \to [0, \infty)$.

1. Lyapunov drift

There exist $\lambda < 1$, $K < \infty$ such that

$$\boxed{QV(x) \le \lambda V(x) + K.}$$

This prevents escape to infinity.

2. Small-set minorization

For $C_R = \{V \le R\}$, there exist $\varepsilon > 0$ and a probability measure ν such that

$$\boxed{Q(x, A) \ge \varepsilon \nu(A), \quad x \in C_R.}$$

This gives uniform mixing inside a compact set.

Harris idea

Drift condition: return to C_R $\quad + \quad$ Minorization: mix inside C_R.

Harris theorem: contraction in weighted total variation

Harris contraction (Hairer and Mattingly, 2008)

Under Lyapunov drift and a small-set minorization conditions there exist $\beta > 0$ and $\bar{\alpha} \in (0, 1)$ such that

$$\boxed{\rho_\beta(\mu Q, \nu Q) \leq \bar{\alpha}\, \rho_\beta(\mu, \nu)}$$

for all probability measures μ, ν with finite V-moment.

Weighted total variation

Let $V : \mathbb{R}^d \to [0, \infty)$ be a Lyapunov function and let $b > 0$. For probability measures μ, ν with finite V-moment, define

$$\rho_b(\mu, \nu) := \int_{\mathbb{R}^d} \big(1 + bV(x)\big)\, |\mu - \nu|(\mathrm{d}x).$$

Interpretation

If $V(x) \to \infty$ as $\|x\| \to \infty$, then

$$\begin{aligned} \text{error near the center} \quad &: \quad (1 + bV_2(x)) \approx 1 \quad \Longrightarrow \quad \text{small cost}, \\ \text{error in the tails} \quad &: \quad (1 + bV_2(x)) \gg 1 \quad \Longrightarrow \quad \text{large cost}. \end{aligned}$$

Why ordinary TV does not capture the geometry of the state space

Consider the toy deterministic $1d$ kernel that brings points back toward the origin:

$$Q(x, \cdot) = \delta_{x/2}(\cdot),$$

Total variation distance

For $x \neq 0$,

$$\|\delta_x - \delta_0\|_{\mathrm{TV}} = 1, \qquad \|\delta_x Q - \delta_0 Q\|_{\mathrm{TV}} = \|\delta_{x/2} - \delta_0\|_{\mathrm{TV}} = 1.$$

So ordinary TV does not see that x moved closer to the center.

Weighted total variation

With $V(x) = \|x\|^2$, for $x \neq 0$, using that $|\delta_x - \delta_0| = \delta_x + \delta_0$,

$$\rho_b(\delta_x, \delta_0) = \int (1 + b\|z\|^2)\, |\delta_x - \delta_0|(\mathrm{d}z) = (1 + b\|x\|^2) + (1 + b\|0\|^2) = 2 + b\|x\|^2.$$

and

$$\rho_b(\delta_x Q, \delta_0 Q) = \rho_b(\delta_{x/2}, \delta_0) = (1 + b\|x/2\|^2) + (1 + b\|0\|^2) = 2 + \frac{b}{4}\|x\|^2.$$

Relationship with TV and Wasserstein

Let $V_2(x) = \|x\|^2$ and $\rho_b(\mu,\nu) := \int_{\mathbb{R}^d}(1 + b\|x\|^2)|\mu-\nu|(\mathrm{d}x)$.

Control of TV and $\mathcal{W}_2$

For probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\boxed{\|\mu-\nu\|_{\mathrm{TV}} \le \frac{1}{2}\rho_b(\mu,\nu), \qquad \mathcal{W}_2^2(\mu,\nu) \le \frac{2}{b}\rho_b(\mu,\nu).}$$

Proof

Since $1 + b\|x\|^2 \ge 1$,

$$\|\mu-\nu\|_{\mathrm{TV}} = \frac{1}{2}|\mu-\nu|(\mathbb{R}^d) \le \frac{1}{2}\rho_b(\mu,\nu).$$

Moreover, using (Villani, 2009, Theorem 6.15)

$$\mathcal{W}_2^2(\mu,\nu) \le 2\int_{\mathbb{R}^d}\|x\|^2|\mu-\nu|(\mathrm{d}x) \le \frac{2}{b}\rho_b(\mu,\nu).$$

Assumptions on the data distribution: Brownian case

Data assumptions

Let $\pi_{\text{data}}(\mathrm{d}x) = p_0(x)\mathrm{d}x$, $p_0 \in C^2(\mathbb{R}^d)$ and assume there exist constants

$$\gamma_0 > 0, \qquad \kappa_0 \geq 0, \qquad C_0 > 0, \qquad m \geq 1,$$

such that, for all $x \in \mathbb{R}^d$,

$$\langle \nabla \log p_0(x), x \rangle \leq -\gamma_0 \|x\|^2 + \kappa_0, \qquad \text{(dissipativity)}$$

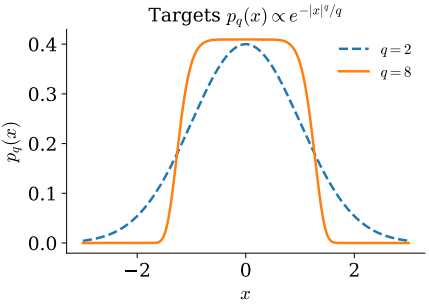
and

$$\|\nabla^2 \log p_0(x)\|_{\mathrm{F}} \leq C_0(1 + \|x\|^m). \qquad \text{(polynomial regularity)}$$

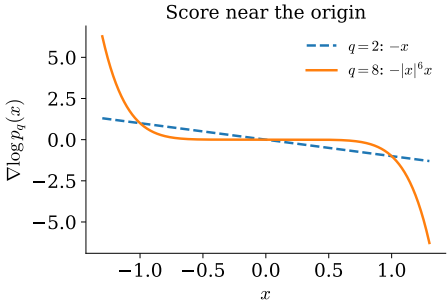
Interpretation: polynomial Jacobian growth

- The condition Jacobian condition is not a Lipschitz assumption, which would require instead $\sup_{x \in \mathbb{R}^d} \|\nabla^2 \log p_0(x)\| < \infty$.
- Here the Jacobian of the score may grow polynomially. This includes strongly confining distribution.

Example: Gaussian vs. strongly confining target



Densities $p_q(x) \propto \exp(-|x|^q/q)$, for $q = 2$ and $q = 8$.



Corresponding scores $\nabla \log p_q(x) = -|x|^{q-2}x$

Compactly supported data

Many data distributions are naturally supported on a bounded set, as density on $\mathbb{R}^d$, this means $p_0(x) = 0$ outside the support, so $\log p_0$ and $\nabla \log p_0$ not well-defined. One may consider

$$p_0^\varepsilon(x) \propto \int \exp\left(-\frac{\|x-y\|^q}{\varepsilon}\right) \pi_{\text{data}}(\mathrm{d}y) \,, \qquad q \geq 2.$$

Dissipativity: inward score and return to the mixing set

Dissipativity means inward radial drift

Outside the ball

$$\|x\|^2 \geq \frac{2\kappa_0}{\gamma_0},$$

the dissipativity condition gives

$$\langle \nabla \log p_0(x), x \rangle \leq -\frac{\gamma_0}{2}\|x\|^2.$$

Equivalently,

$$\left\langle \nabla \log p_0(x), \frac{x}{\|x\|} \right\rangle \leq -\frac{\gamma_0}{2}\|x\|.$$

Thus, far from the origin, the score points back to the center.

Why this matters for Harris

The reverse dynamics is driven by the score. The dissipativity condition prevents the process from escaping to infinity and pushes it back toward a region where minorization can produce mixing.

Dissipativity implies Gaussian-type tail control

Fix a direction $u \in \mathbb{S}^{d-1}$ and write $x = ru$. Then

$$\frac{\mathrm{d}}{\mathrm{d}r} \log p_0(ru) = \frac{\langle \nabla \log p_0(ru), ru \rangle}{r}.$$

By dissipativity,

$$\frac{\mathrm{d}}{\mathrm{d}r} \log p_0(ru) \leq -\gamma_0 r + \frac{\kappa_0}{r}.$$

For $r \geq \sqrt{2\kappa_0/\gamma_0}$, this gives

$$\frac{\mathrm{d}}{\mathrm{d}r} \log p_0(ru) \leq -\frac{\gamma_0}{2} r.$$

Integrating the above

$$\log p_0(ru) \leq C - \frac{\gamma_0}{4} r^2,$$

and therefore

$$p_0(x) \leq C \exp\left(-\frac{\gamma_0}{4}\|x\|^2\right) \qquad \text{for large } \|x\|.$$

Message

Dissipativity means enforce Gaussian-type tail decay.

Stability under Gaussian perturbation

Dissipativity propagation through Gaussian smoothing

$$\langle \nabla \log p_0(x), x \rangle \leq -\gamma_0 \|x\|^2 + \kappa_0 \quad \Longrightarrow \quad \langle \nabla \log p_t(x), x \rangle \leq -\gamma_t \|x\|^2 + \kappa_t \,, \text{for all } t \geq 0$$

The smoothed score remains dissipative under Gaussian perturbation and, where, in the Brownian convention $p_t = p_0 * \mathcal{N}(0, 2t\mathrm{I}_d)$, one can take

$$\gamma_t = \frac{\gamma_0}{1 + 4\gamma_0 t}, \qquad \kappa_t = \frac{\kappa_0 + d}{1 + 4\gamma_0 t}.$$

⚠ the estimates are quantitative !

Remark: polynomial regularity is also stable

The polynomial growth control on the score Jacobian is also propagated by the Gaussian perturbation.

⚠ This transfers assumptions made on π_{data} to the smoothed marginals of the forward and therefore to the score function driving the reverse dynamics.

Ingredient 1: From dissipativity to a Lyapunov drift bound

The reverse process has generator, at reverse time u,

$$\mathcal{A}_u f(x) = 2\,\langle \nabla \log p_{T-u}(x), \nabla f(x)\rangle + \Delta f(x).$$

Apply the generator to $V_2(x) = \|x\|^2$

Since $\nabla V_2(x) = 2x$ and $\Delta V_2(x) = 2d$, we get

$$\mathcal{A}_u V_2(x) = 4\langle \nabla \log p_{T-u}(x), x\rangle + 2d.$$

Using the propagated dissipativity, $\mathcal{A}_u V_2(x) \le -4\gamma_{T-u} V_2(x) + (4\kappa_{T-u} + 2d) \le -\widetilde{\gamma}_u V_2 + \widetilde{\kappa}_u$.

Semigroup Lyapunov drift

By Dynkin's formula and Grönwall,

$$Q_{t|s} V_2(x) \le \lambda_{t|s} V_2(x) + K_{t|s}\,,$$

with

$$\lambda_{t|s} = \exp\left(-\int_s^t \widetilde{\gamma}_u \,\mathrm{d}u\right), \quad \text{and} \quad K_{t|s} = \int_s^t \exp\left(-\int_u^t \widetilde{\gamma}_v \,\mathrm{d}v\right) \widetilde{\kappa}_u \,\mathrm{d}u.$$

Ingredient 2: Local minorization

The reverse kernel admits the Bayes representation

$$Q_{t|s}(x, \mathrm{d}y) = \frac{\varphi_{2(t-s)}(x-y)\, p_{T-t}(y)}{p_{T-s}(x)}\, \mathrm{d}y, \qquad 0 \le s < t \le T,$$

where φ_a denotes the density of $\mathcal{N}(0, a\mathrm{I}_d)$.

Restriction to a compact set

Fix

$$C_r = \{x \in \mathbb{R}^d : \|x\| \le r\}.$$

For $x \in C_r$, the Gaussian factor admits a common lower envelope:

$$\varphi_{2(t-s)}(x-y) \ge c_{r,t-s}\, \varphi_{t-s}(y), \qquad c_{r,t-s} > 0.$$

Moreover, since p_{T-s} is Gaussian-smoothed, $p_{T-s}(x) \le M_{T-s} < \infty$.

Localized Doeblin condition

Therefore, there exist a probability measure $\nu_{t|s}$ and $\varepsilon^{(r)}_{t|s} > 0$ such that, for all $x \in C_r$,

$$Q_{t|s}(x, A) \ge \varepsilon^{(r)}_{t|s}\, \nu_{t|s}(A), \qquad A \in \mathcal{B}(\mathbb{R}^d).$$

Conclusion: Harris forgetting of the reverse kernel

Two ingredients

For $V_2(x) = \|x\|^2$, for $0 \leq s < t \leq T$, the previous estimates give:

$$Q_{t|s} V_2(x) \leq \lambda^{(2)}_{t|s} V_2(x) + K^{(2)}_{t|s}, \qquad \lambda^{(2)}_{t|s} < 1,$$

and, on $C_r = \{x : \|x\| \leq r\}$,

$$Q_{t|s}(x, \cdot) \geq \varepsilon^{(r)}_{t|s} \nu_{t|s}(\cdot), \qquad x \in C_r.$$

Harris contraction

For any, $r^2 > \frac{2K^{(2)}_{t|s}}{1-\lambda^{(2)}_{t|s}}$, $a_0 \in (0, \varepsilon^{(r)}_{t|s})$, $\eta_0 \in \left(\lambda^{(2)}_{t|s} + \frac{2K^{(2)}_{t|s}}{r^2}, 1\right)$, $b^{(r)}_{s,t} := \frac{a_0}{K^{(2)}_{t|s}}$. Then

$$\rho_{b^{(r)}_{s,t}}\left(\mu Q_{t|s}, \nu Q_{t|s}\right) \leq \bar{\alpha}_{t|s} \rho_{b^{(r)}_{s,t}}(\mu, \nu),$$

with

$$\bar{\alpha}_{t|s} = \left[1 - \left(\varepsilon^{(r)}_{t|s} - a_0\right)\right] \vee \frac{2 + r^2 b^{(r)}_{s,t} \eta_0}{2 + r^2 b^{(r)}_{s,t}} < 1.$$

Stability bound

SGMs in practice: three approximations

$\boxed{p_T Q_{0:T} = \pi_{\text{data}}}$ **is almost generative, but . . .**

1. Initialization

The exact terminal law still depends on the data:

$$\overrightarrow{X}_T \overset{\mathcal{L}}{=} \overrightarrow{X}_0 + \sqrt{2T}\, Z.$$

Equivalently,

$$p_T = p_0 * \mathcal{N}(0, 2T\mathrm{I}_d).$$

For large T,

$$p_T \approx \pi_\infty = N(0, 2T\mathrm{I}_d).$$

Thus, $p_T Q_{0:T} \approx \pi_\infty Q_{0:T}$.

2. Score approximation

The reverse drift uses the unknown score

$$\nabla \log p_{T-t}(x).$$

In practice, learn

$$\mathsf{s}_\theta(x, T-t) \approx \nabla \log p_{T-t}(x).$$

Hence

$$\pi_{\text{data}} \approx \pi_\infty Q^\theta_{0:T}.$$

3. Discretization

Choose a grid

$$t_k = kh, \qquad h = \frac{T}{N}.$$

Replace the continuous learned reverse process by Euler–Maruyama:

$$\widehat{\pi}^\theta_N = \pi_\infty \bar{Q}^\theta_0 \cdots \bar{Q}^\theta_{N-1}.$$

$$\boxed{\bar{X}^{\theta,N}_{k+1} = \bar{X}^{\theta,N}_k + 2h\,\mathsf{s}_\theta(\bar{X}^{\theta,N}_k, T-t_k) + \sqrt{2h}\, Z_{k+1}\,, \quad \bar{X}^{\theta,N}_0 \sim \pi_\infty\,, \qquad Z_{k+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathrm{I}_d).}$$

Generated distribution and target error

Let $t_k = kh$, $h = T/N$. The exact VE/Brownian reverse process satisfies

$$\mathrm{d}\overleftarrow{X}_t = 2\nabla \log p_{T-t}(\overleftarrow{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t.$$

Kernels

For a bounded test function f, define

$$Q_k f(x) := \mathbb{E}\left[f(\overleftarrow{X}_{t_{k+1}}) \,\middle|\, \overleftarrow{X}_{t_k} = x \right], \qquad k = 0, \dots, N-1.$$

and

$$\bar{Q}_k^\theta f(x) := \mathbb{E}\left[f\left(\bar{X}_{k+1}^{\theta,N}\right) \,\middle|\, \bar{X}_k^{\theta,N} = x \right], \qquad k = 0, \dots, N-1.$$

Thus the generated law is

$$\widehat{\pi}_{\infty,N}^\theta = \pi_\infty \bar{Q}_0^\theta \cdots \bar{Q}_{N-1}^\theta$$

Stability objective

We want to control some discrepancies on proba. measures ρ

$$\rho\left(\pi_{\mathrm{data}}, \widehat{\pi}_{\infty,N}^\theta\right)$$

Forgetting and stability analysis

The stability question

If the reverse kernel contracts: $\rho(\mu Q_k, \nu Q_k) \leq \alpha_k \rho(\mu, \nu), \qquad \alpha_k < 1.$

Define the ideal laws at step k, a marginal of the reverse process given by

$$\mu_{k+1} = p_T Q_{t_k|0},$$

and $\widehat{\mu}_k = \pi_\infty \bar{Q}_0^\theta \cdots \bar{Q}_{k-1}^\theta$ is the practical approximation of μ_k.

The *accumulated error at time k*, e_k writes as

$$\begin{aligned} e_{k+1} &= \rho(\mu_k Q_k, \widehat{\mu}_k \bar{Q}_k^\theta) \\ &\leq \rho(\mu_k Q_k, \widehat{\mu}_k Q_k) + \rho(\widehat{\mu}_k Q_k, \widehat{\mu}_k \bar{Q}_k^\theta) \\ &\leq \alpha_k e_k + \delta_k. \end{aligned}$$

where δ_k is a local error and α_k measures how much the remaining dynamics remembers the past.

Forgetting mechanism

$$e_N \leq \left(\prod_{\ell=0}^{N-1} \alpha_\ell \right) e_0 + \sum_{k=0}^{N-1} \left(\prod_{\ell=k+1}^{N-1} \alpha_\ell \right) \delta_k.$$

Existing stability approaches: Girsanov-based techniques

Compare the exact and approximated reverse SDE

$$\mathrm{d}\overleftarrow{X}_t = 2\nabla \log p_{T-t}(\overleftarrow{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \qquad \mathrm{d}X^\theta_t = 2\mathsf{s}_\theta(X^\theta_t, T-t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,.$$

Path-space comparison

Let $\mathbb{P}^{\mu}_{[0,T]}$ and $\mathbb{P}^{\theta,\mu}_{[0,T]}$ denote the path laws of these two processes started from μ. By Girsanov's theorem,

$$\mathsf{KL}\left(\mathbb{P}^{\mu}_{[0,T]} \,\middle\|\, \mathbb{P}^{\theta,\mu}_{[0,T]}\right) \le \int_0^T \mathbb{E}_{\mathbb{P}^\mu}\left[\left\|\nabla \log p_{T-t}(\overleftarrow{X}_t) - \mathsf{s}_\theta(\overleftarrow{X}_t, T-t)\right\|^2\right] \mathrm{d}t.$$

Terminal-time bound by data processing

$$\mathsf{KL}\left(\mathcal{L}(\overleftarrow{X}_T) \,\middle\|\, \mathcal{L}(X^\theta_T)\right) \le \mathsf{KL}\left(\mathbb{P}^{\mu}_{[0,T]} \,\middle\|\, \mathbb{P}^{\theta,\mu}_{[0,T]}\right).$$

Girsanov bound: discrete reading

Split the time interval with $t_k = kh$, $h = T/N$. The Girsanov bound can be written as

$$\mathrm{KL}\Big(\mathcal{L}(\overleftarrow{X}_T)\,\Big\|\,\mathcal{L}(X_T^\theta)\Big) \le \mathrm{KL}(\pi_{\mathrm{data}} \,\|\, \pi_\infty) + \sum_{k=0}^{N-1} \Delta_k^{\mathrm{score}},$$

where $\Delta_k^{\mathrm{score}} := \int_{t_k}^{t_{k+1}} \mathbb{E}_{\mathbb{P}^\mu}\left[\left\|\nabla \log p_{T-t}(\overleftarrow{X}_t) - \mathsf{s}_\theta(\overleftarrow{X}_t, T-t)\right\|^2\right] \mathrm{d}t.$

Accumulation

$$e_N \lesssim e_0 + \sum_{k=0}^{N-1} \delta_k.$$

Every local error is paid the same.

Forgetting

One proves

$$e_{k+1} \le \bar{\alpha} e_k + \delta_k, \qquad \bar{\alpha} < 1.$$

After iteration,

$$e_N \lesssim \bar{\alpha}^N e_0 + \sum_{k=0}^{N-1} \bar{\alpha}^{N-1-k} \delta_k.$$

Old errors are damped.

Local one-step kernel error

At step k, compare the exact reverse kernel Q_k with the practical Euler kernel $\bar{Q}_k^\theta$:

$$\delta_k := \rho_b\left(\widehat{\mu}_k Q_k, \widehat{\mu}_k \bar{Q}_k^\theta\right).$$

Split the local error

Introduce $\bar{Q}_k$, the Euler kernel using the *exact* score. Then

$$\delta_k \le \underbrace{\rho_b\left(\widehat{\mu}_k Q_k, \widehat{\mu}_k \bar{Q}_k\right)}_{\text{local discretization error}} + \underbrace{\rho_b\left(\widehat{\mu}_k \bar{Q}_k, \widehat{\mu}_k \bar{Q}_k^\theta\right)}_{\text{local score error}}.$$

Schematic estimates

For the uniform grid $h = T/N$, one obtains

$$\rho_b\left(\widehat{\mu}_k Q_k, \widehat{\mu}_k \bar{Q}_k\right) \lesssim C_k^{\mathrm{disc}} h,$$

and, with $E_k(x) := \nabla \log p_{T-t_k}(x) - \mathsf{s}_\theta(x, T - t_k)$,

$$\rho_b\left(\widehat{\mu}_k \bar{Q}_k, \widehat{\mu}_k \bar{Q}_k^\theta\right) \lesssim C_k^{\mathrm{net}} \sqrt{h}\, \|E_k\|_{L^2(\widehat{\mu}_k)}.$$

A forgetting stability bound

One-step estimate

For the uniform grid $h = T/N$, the local error satisfies schematically

$$\boxed{\delta_k \lesssim C_k^{\mathrm{disc}} h + C_k^{\mathrm{net}} \sqrt{h} \, \|E_k\|_{L^2(\widehat{\mu}_k)}.}$$

Final bound

Injecting this into the discounted stability estimate gives

$$\rho_b\left(\pi_{\mathrm{data}}, \widehat{\pi}_N^{\theta}\right) \lesssim \bar{\alpha}^N \rho_b\left(p_T, \pi_\infty\right) + \sum_{k=0}^{N-1} \bar{\alpha}^{N-1-k} \left[C_k^{\mathrm{disc}} h + C_k^{\mathrm{net}} \sqrt{h} \, \|E_k\|_{L^2(\widehat{\mu}_k)} \right].$$

Numerical illustration

Numerical illustration: what is tested?

Goal

The experiments are designed to isolate the forgetting mechanism:

$$\text{perturb early in the reverse trajectory} \implies \text{small final effect,}$$

whereas

$$\text{perturb late in the reverse trajectory} \implies \text{large final effect.}$$

Two perturbation protocols

1. **Initialization perturbation:** perturb the cloud at an intermediate noise level, then run the reverse sampler to the data time.
2. **Local score perturbation:** keep the initialization fixed, but perturb the score at one single reverse step.

⚠ When the score is known and there is no discretization error our theory gives a precise estimate of such phenomenon for the *weigthed TV distance*.

Two controlled perturbation experiments

Let t_{pert} denote the forward noise level where the perturbation is introduced.

1. Initialization perturbation

1. Sample particles from the forward marginal $p_{t_{\mathrm{pert}}}$.
2. Shift the cloud in a fixed direction:
$$\vec{X}^{\mathrm{pert}}_{t_{\mathrm{pert}}} = \vec{X}_{t_{\mathrm{pert}}} + \lambda u.$$
3. Run the reverse sampler back to time 0.
4. Compare the final law with π_{data}.

2. Local score perturbation

1. Start from the same terminal initialization.
2. Run the exact-score reverse sampler.
3. At one step k_{pert}, replace
$$\nabla \log p_{T-t_k} \quad \text{by} \quad \nabla \log p_{T-t_k} + \lambda u.$$
4. Complete the reverse sampler and measure the final error.

Expected signature of forgetting

The final error should be smaller when t_{pert} is far from the data time, because more reverse dynamics remains after the perturbation.

Gaussian targets: three controlled geometries

For a Gaussian target $\pi_{\text{data}} = \mathcal{N}(0, \Sigma)$ and VE noising,

$$\overrightarrow{X}_t = X_0 + \sqrt{2t}\, Z, \qquad p_t = \mathcal{N}(0, \Sigma + 2t\mathrm{I}_d).$$

Hence the score is explicit: $\nabla \log p_t(x) = -(\Sigma + 2t\mathrm{I}_d)^{-1} x$.

Isotropic

$$\Sigma^{(\text{iso})} = \sigma^2 \mathrm{I}_d, \qquad \sigma^2 = 0.1.$$

All directions have the same variance.

Anisotropic

$$\Sigma^{(\text{het})} = \operatorname{diag}(v_1, \ldots, v_d),$$

$$v_j = \begin{cases} 1, & 1 \le j \le 5, \\ 10^{-3}, & 6 \le j \le d. \end{cases}$$

Different directions have different scales.

Correlated

$$\Sigma^{(\text{corr})}_{jj} = 1,$$

$$\Sigma^{(\text{corr})}_{jj'} = \frac{1}{\sqrt{|j - j'| + 1}}, \quad j \neq j'.$$

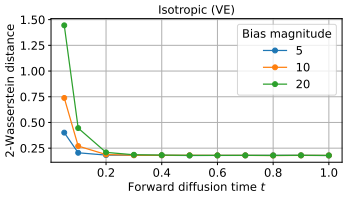
Coordinates are strongly correlated.

Why Gaussian experiments?

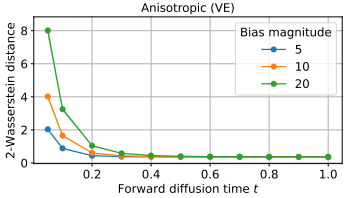
They isolate the forgetting phenomenon in a setting where the score and the reverse dynamics are explicit.

Gaussian targets: perturbation sensitivity

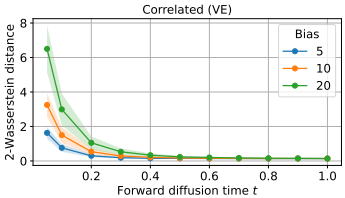
Initialization perturbation



Isotropic

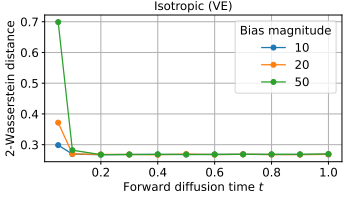


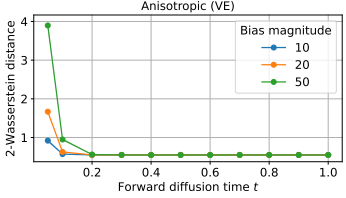
Anisotropic

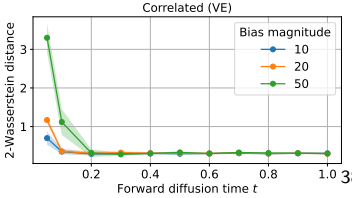


Correlated

One-step score perturbation







Gaussian mixture experiment

Controlled multimodal target

We consider a Gaussian mixture model in dimension $d = 50$, with 25 components. The component means are arranged on a 5×5 grid in the first two coordinates, so that samples can be visualized by projecting onto these two coordinates.

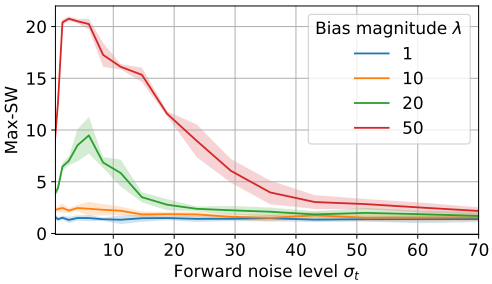
Two perturbation protocols

initialization perturbation and one-step score perturbation.

For each perturbation time and perturbation magnitude λ, we run the reverse sampler and compare the final output with π_{data}.

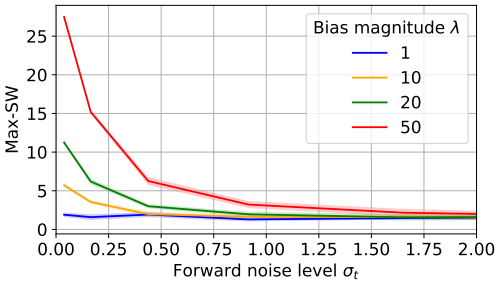
GMM: perturbation sensitivity

Initialization perturbation



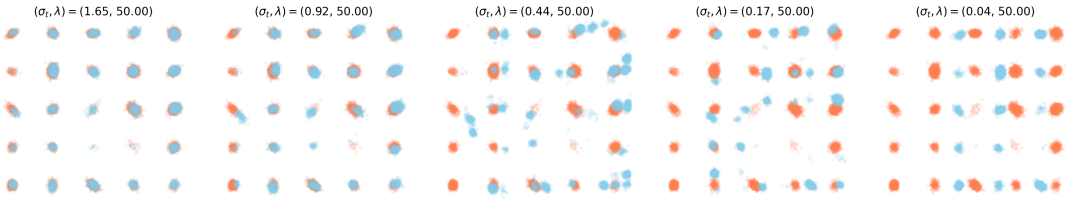
Perturb the cloud at a chosen noise level, then reverse.

One-step score perturbation



Perturb the score at one reverse step, then continue.

GMM: visualizing one-step score perturbations



Visual message

As the perturbation is made closer to the data time, the final samples deviate more visibly from the target mixture. Late score errors have little time left to be forgotten.

CIFAR-10: one-step score perturbation

Real-data perturbation experiment

Using a pretrained EDM VP model on CIFAR-10, perturb the denoiser once.

Step	0	25	50	70	75	80	85	90	95
FID	13.3	13.0	13.1	13.6	14.4	16.4	28.3	153	304
Max-SW	0.011	0.014	0.016	0.033	0.044	0.060	0.094	0.266	0.779

References

L. Ambrogioni. The statistical thermodynamics of generative diffusion models: Phase transitions, symmetry breaking and critical instability. 2024.

G. Biroli, T. Bonnaire, V. De Bortoli, and M. Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 2024.

M. Gentiloni-Silveri and A. Ocello. Beyond log-concavity and score regularity: Improved convergence bounds for score-based generative models in w2-distance. In *Forty-second International Conference on Machine Learning*, 2025.

M. Hairer and J. C. Mattingly. Yet another look at harris' ergodic theorem for markov chains. *arXiv preprint arXiv:0810.2777*, 2008. URL `https://arxiv.org/abs/0810.2777`.

U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14(4):1188–1205, 1986.

C. Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9.